Proceedings of the 40th Annual Meeting of the Association for
Computational Linguistics (ACL), Philadelphia, July 2002, pp. 360-367.

# Bootstrapping

Steven Abney
AT&T Laboratories – Research
180 Park Avenue
Florham Park, NJ, USA, 07932

## Abstract

This paper refines the analysis of co-training, defines and evaluates a new co-training algorithm that has theoretical justification, gives a theoretical justification for the Yarowsky algorithm, and shows that co-training and the Yarowsky algorithm are based on different independence assumptions.

## 1 Overview

The term *bootstrapping* here refers to a problem setting in which one is given a small set of labeled data and a large set of unlabeled data, and the task is to induce a classifier. The plenitude of unlabeled natural language data, and the paucity of labeled data, have made bootstrapping a topic of interest in computational linguistics. Current work has been spurred by two papers, (Yarowsky, 1995) and (Blum and Mitchell, 1998).

Blum and Mitchell propose a conditional independence assumption to account for the efficacy of their algorithm, called *co-training*, and they give a proof based on that conditional independence assumption. They also give an intuitive explanation of why co-training works, in terms of maximizing agreement on unlabeled data between classifiers based on different "views" of the data. Finally, they suggest that the Yarowsky algorithm is a special case of the co-training algorithm.

The Blum and Mitchell paper has been very influential, but it has some shortcomings. The proof they give does not actually apply directly to the co-training algorithm, nor does it directly justify the intuitive account in terms of classifier agreement on unlabeled data, nor, for that matter, does the co-training algorithm directly seek to find classifiers that agree on unlabeled data. Moreover, the suggestion that the Yarowsky algorithm is a special case of co-training is based on an incidental detail of the particular application that Yarowsky considers, not on the properties of the core algorithm.

In recent work, (Dasgupta et al., 2001) prove that a classifier has low generalization error if it agrees on unlabeled data with a second classifier based on a different "view" of the data. This addresses one of the shortcomings of the original co-training paper: it gives a proof that justifies searching for classifiers that agree on unlabeled data.

I extend this work in two ways. First, (Dasgupta et al., 2001) assume the same conditional independence assumption as proposed by Blum and Mitchell. I show that that independence assumption is remarkably powerful, and violated in the data; however, I show that a weaker assumption suffices. Second, I give an algorithm that finds classifiers that agree on unlabeled data, and I report on an implementation and empirical results.

Finally, I consider the question of the relation between the co-training algorithm and the Yarowsky algorithm. I suggest that the Yarowsky algorithm is actually based on a different independence assumption, and I show that, if the independence assumption holds, the Yarowsky algorithm is effective at finding a high-precision classifier.

## 2 Problem Setting and Notation

A bootstrapping problem consists of a space of instances $\mathcal{X}$, a set of labels $\mathcal{L}$, a function

$Y : \mathcal{X} \rightarrow \mathcal{L}$ assigning labels to instances, and a space of rules mapping instances to labels. Rules may be partial functions; we write $F(x) = \perp$ if $F$ abstains (that is, makes no prediction) on input $x$. "Classifier" is synonymous with "rule".

It is often useful to think of rules and labels as sets of instances. A binary rule $F$ can be thought of as the characteristic function of the set of instances $\{x : F(x) = +\}$. Multi-class rules also define useful sets when a particular target class $\ell$ is understood. For any rule $F$, we write $F_\ell$ for the set of instances $\{x : F(x) = \ell\}$, or (ambiguously) for that set's characteristic function.

We write $\bar{F}_\ell$ for the complement of $F_\ell$, either as a set or characteristic function. Note that $\bar{F}_\ell$ contains instances on which $F$ abstains. We write $F_{\bar{\ell}}$ for $\{x : F(x) \neq \ell \wedge F(x) \neq \perp\}$. When $F$ does not abstain, $\bar{F}_\ell$ and $F_{\bar{\ell}}$ are identical.

Finally, in expressions like $\Pr[F = +|Y = +]$ (with square brackets and "Pr"), the functions $F(x)$ and $Y(x)$ are used as random variables. By contrast, in the expression $P(F|Y)$ (with parentheses and "$P$"), $F$ is the set of instances for which $F(x) = +$, and $Y$ is the set of instances for which $Y(x) = +$.

## 3 View Independence

Blum and Mitchell assume that each instance $x$ consists of two "views" $x_1, x_2$. We can take this as the assumption of functions $X_1$ and $X_2$ such that $X_1(x) = x_1$ and $X_2(x) = x_2$. They propose that views are conditionally independent given the label.

**Definition 1** *A pair of views $x_1$, $x_2$ satisfy* view independence *just in case:*

$$\begin{aligned} \Pr[X_1 = x_1|X_2 = x_2, Y = y] &= \Pr[X_1 = x_1|Y = y] \\ \Pr[X_2 = x_2|X_1 = x_1, Y = y] &= \Pr[X_2 = x_2|Y = y] \end{aligned}$$

*A classification problem instance satisfies view independence just in case all pairs $x_1$, $x_2$ satisfy view independence.*

There is a related independence assumption that will prove useful. Let us define $\mathcal{H}_1$ to consist of rules that are functions of $X_1$ only, and define $\mathcal{H}_2$ to consist of rules that are functions of $X_2$ only.

**Definition 2** *A pair of rules $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ satisfies* rule independence *just in case, for all $u, v, y$:*

$$\Pr[F = u|G = v, Y = y] \quad = \quad \Pr[F = u|Y = y]$$

*and similarly for $F \in \mathcal{H}_2$, $G \in \mathcal{H}_1$. A classification problem instance satisfies rule independence just in case all opposing-view rule pairs satisfy rule independence.*

If instead of generating $\mathcal{H}_1$ and $\mathcal{H}_2$ from $X_1$ and $X_2$, we assume a set of features $\mathcal{F}$ (which can be thought of as binary rules), and take $\mathcal{H}_1 = \mathcal{H}_2 = \mathcal{F}$, rule independence reduces to the Naive Bayes independence assumption.

The following theorem is not difficult to prove; we omit the proof.

**Theorem 1** *View independence implies rule independence.*

## 4 Rule Independence and Bootstrapping

Blum and Mitchell's paper suggests that rules that agree on unlabelled instances are useful in bootstrapping.

**Definition 3** *The* agreement rate *between rules $F$ and $G$ is*

$$\Pr[F = G|F, G \neq \perp]$$

Note that the agreement rate between rules makes no reference to labels; it can be determined from unlabeled data.

The algorithm that Blum and Mitchell describe does not explicitly search for rules with good agreement; nor does agreement rate play any direct role in the learnability proof given in the Blum and Mitchell paper.

The second lack is emended in (Dasgupta et al., 2001). They show that, if view independence is satisfied, then the agreement rate between opposing-view rules $F$ and $G$ upper bounds the error of $F$ (or $G$). The following statement of the theorem is simplified and assumes non-abstaining binary rules.

**Theorem 2** *For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy rule independence and are nontrivial predictors in the sense that* $\min_u \Pr[F = u] > \Pr[F \neq$

$G$], *one of the following inequalities holds:*

$$\begin{array}{rcl} \Pr[F \neq Y] & \leq & \Pr[F \neq G] \\ \Pr[\bar{F} \neq Y] & \leq & \Pr[F \neq G] \end{array}$$

If $F$ agrees with $G$ on all but $\epsilon$ unlabelled instances, then either $F$ or $\bar{F}$ predicts $Y$ with error no greater than $\epsilon$. A small amount of labelled data suffices to choose between $F$ and $\bar{F}$.

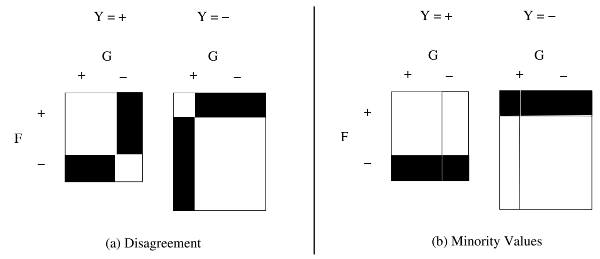
I give a geometric proof sketch here; the reader is referred to the original paper for a formal proof. Consider figures 1 and 2. In these diagrams, area represents probability. For example, the leftmost box (in either diagram) represents the instances for which $Y = +$, and the area of its upper left quadrant represents $\Pr[F = +, G = +, Y = +]$. Typically, in such a diagram, either the horizontal or vertical line is broken, as in figure 2. In the special case in which rule independence is satisfied, both horizontal and vertical lines are unbroken, as in figure 1.

Theorem 2 states that disagreement upper bounds error. First let us consider a lemma, to wit: disagreement upper bounds minority probabilities. Define the *minority value* of $F$ given $Y = y$ to be the value $u$ with least probability $\Pr[F = u|Y = y]$; the *minority probability* is the probability of the minority value. (Note that minority probabilities are *conditional* probabilities, and distinct from the marginal probability $\min_u \Pr[F = u]$ mentioned in the theorem.)

In figure 1a, the areas of disagreement are the upper right and lower left quadrants of each box, as marked. The areas of minority values are marked in figure 1b. It should be obvious that the area of disagreement upper bounds the area of minority values.

The *error values* of $F$ are the values opposite to the values of $Y$: the error value is $-$ when $Y = +$ and $+$ when $Y = -$. When minority values are error values, as in figure 1, disagreement upper bounds error, and theorem 2 follows immediately.

However, three other cases are possible. One possibility is that minority values are opposite to error values. In this case, the minority values of $\bar{F}$ are error values, and disagreement between $F$ and $G$ upper bounds the error of $\bar{F}$.



(a) Disagreement

(b) Minority Values

Figure 1: Disagreement upper-bounds minority probabilities.

This case is admitted by theorem 2. In the final two cases, minority values are the same regardless of the value of $Y$. In these cases, however, the predictors do not satisfy the "nontriviality" condition of theorem 2, which requires that $\min_u \Pr[F = u]$ be greater than the disagreement between $F$ and $G$.

## 5 The Unreasonableness of Rule Independence

Rule independence is a very strong assumption; one remarkable consequence will show just how strong it is. The *precision* of a rule $F$ is defined to be $\Pr[Y = +|F = +]$. (We continue to assume non-abstaining binary rules.) If rule independence holds, knowing the precision of any one rule allows one to *exactly compute* the precision of every other rule given only *unlabeled* data and knowledge of the size of the target concept.

Let $F$ and $G$ be arbitrary rules based on independent views. We first derive an expression for the precision of $F$ in terms of $G$. Note that the second line is derived from the first by rule independence.

$$\begin{array}{rcl} P(FG) & = & P(F|GY)P(GY) + P(F|G\bar{Y})P(G\bar{Y}) \\ & = & P(F|Y)P(GY) + P(F|\bar{Y})P(G\bar{Y}) \\ P(G|F) & = & P(Y|F)P(G|Y) + [1 - P(Y|F)]P(G|\bar{Y}) \\ P(Y|F) & = & \frac{P(G|F) - P(G|\bar{Y})}{P(G|Y) - P(G|\bar{Y})} \end{array}$$

To compute the expression on the righthand side of the last line, we require $P(Y|G)$, $P(Y)$, $P(G|F)$, and $P(G)$. The first value, the precision of $G$, is assumed known. The second value, $P(Y)$, is also assumed known; it can at any rate be estimated from a small amount of labeled data. The last two values, $P(G|F)$ and $P(G)$, can be computed from unlabeled data.

Thus, given the precision of an arbitrary rule $G$, we can compute the precision of any other-view rule $F$. Then we can compute the precision of rules based on the same view as $G$ by using the precision of some other-view rule $F$. Hence we can compute the precision of every rule given the precision of any one.

## 6 Some Data

The empirical investigations described here and below use the data set of (Collins and Singer, 1999). The task is to classify names in text as person, location, or organization. There is an unlabeled training set containing 89,305 instances, and a labeled test set containing 289 persons, 186 locations, 402 organizations, and 123 "other", for a total of 1,000 instances.

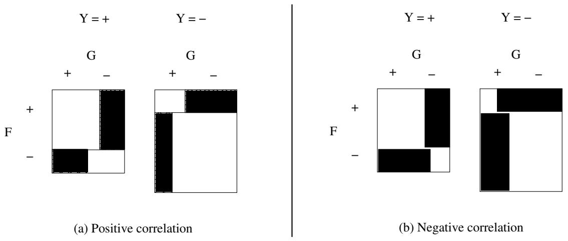
Instances are represented as lists of features. *Intrinsic features* are the words making up the name, and *contextual features* are features of the syntactic context in which the name occurs. For example, consider *Bruce Kaplan, president of Metals Inc.* This text snippet contains two instances. The first has intrinsic features `N:Bruce-Kaplan`, `C:Bruce`, and `C:Kaplan` ("N" for the complete name, "C" for "contains"), and contextual feature `M:president` ("M" for "modified by"). The second instance has intrinsic features `N:Metals-Inc`, `C:Metals`, `C:Inc`, and contextual feature `X:president-of` ("X" for "in the context of").

Let us define $Y(x) = +$ if $x$ is a "location" instance, and $Y(x) = -$ otherwise. We can estimate $P(Y)$ from the test sample; it contains 186/1000 location instances, giving $P(Y) = .186$.

Let us treat each feature $F$ as a rule predicting $+$ when $F$ is present and $-$ otherwise. The precision of $F$ is $P(Y|F)$. The internal feature `N:New-York` has precision 1. This permits us to compute the precision of various contextual features, as shown in the "Co-training" column of Table 1. We note that the numbers do not even look like probabilities. The cause is the failure of view independence to hold in the data, combined with the instability of the estimator. (The "Yarowsky" column uses a seed rule to estimate $P(Y|F)$, as is done in the Yarowsky algorithm, and the "Truth" column shows the true value of $P(Y|F)$.)

| $F$ | Co-training | Yarowsky | Truth |
|---|---|---|---|
| `M:chairman` | -12.7 | .068 | .030 |
| `X:Company-of` | 10.2 | .979 | .989 |
| `X:court-in` | -.183 | 1.00 | .981 |
| `X:Company-in` | 75.7 | 1.00 | .986 |
| `X:firm-in` | -9.94 | .952 | .949 |
| `X:%-in` | -15.2 | .875 | .192 |
| `X:meeting-in` | -2.25 | 1.00 | .753 |

Table 1: Some data



(a) Positive correlation

(b) Negative correlation

Figure 2: Deviation from conditional independence.

## 7 Relaxing the Assumption

Nonetheless, the unreasonableness of view independence does not mean we must abandon theorem 2. In this section, we introduce a weaker assumption, one that *is* satisfied by the data, and we show that theorem 2 holds under this weaker assumption.

There are two ways in which the data can diverge from conditional independence: the rules may either be positively or negatively correlated, given the class value. Figure 2a illustrates positive correlation, and figure 2b illustrates negative correlation.

If the rules are negatively correlated, then their disagreement (shaded in figure 2) is larger than if they are conditionally independent, and the conclusion of theorem 2 is maintained a fortiori. Unfortunately, in the data, they are positively correlated, so the theorem does not apply.

Let us quantify the amount of deviation from conditional independence. We define the *conditional dependence* of $F$ and $G$ given $Y = y$ to be $d_y =$

$$\frac{1}{2}\sum_{u,v} |\Pr[G = v|Y = y, F = u] - \Pr[G = v|Y = y]|$$

If $F$ and $G$ are conditionally independent, then $d_y = 0$.

This permits us to state a weaker version of rule independence:

**Definition 4** *Rules $F$ and $G$ satisfy* weak rule dependence *just in case, for $y \in \{+, -\}$:*

$$d_y \leq p_2 \frac{q_1 - p_1}{2p_1 q_1}$$

*where $p_1 = \min_u \Pr[F = u|Y = y]$, $p_2 = \min_u \Pr[G = u|Y = y]$, and $q_1 = 1 - p_1$.*

By definition, $p_1$ and $p_2$ cannot exceed 0.5. If $p_1 = 0.5$, then weak rule dependence reduces to independence: if $p_1 = 0.5$ and weak rule dependence is satisfied, then $d_y$ must be 0, which is to say, $F$ and $G$ must be conditionally independent. However, as $p_1$ decreases, the permissible amount of conditional dependence increases.

We can now state a generalized version of theorem 2:

**Theorem 3** *For all $F \in \mathcal{H}_1$, $G \in \mathcal{H}_2$ that satisfy weak rule dependence and are nontrivial predictors in the sense that $\min_u \Pr[F = u] > \Pr[F \neq G]$, one of the following inequalities holds:*

$$\begin{array}{rcl} \Pr[F \neq Y] & \leq & \Pr[F \neq G] \\ \Pr[\bar{F} \neq Y] & \leq & \Pr[F \neq G] \end{array}$$

Consider figure 3. This illustrates the most relevant case, in which $F$ and $G$ are positively correlated given $Y$. (Only the case $Y = +$ is shown; the case $Y = -$ is similar.) We assume that the minority values of $F$ are error values; the other cases are handled as in the discussion of theorem 2.

Let $u$ be the minority value of $G$ when $Y = +$. In figure 3, $a$ is the probability that $G = u$ when $F$ takes its minority value, and $b$ is the probability that $G = u$ when $F$ takes its majority value.

The value $r = a - b$ is the difference. Note that $r = 0$ if $F$ and $G$ are conditionally independent given $Y = +$. In fact, we can show that $r$ is exactly our measure $d_y$ of conditional dependence:

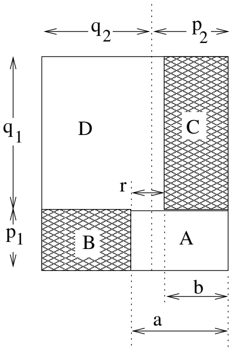


Figure 3: Positive correlation, $Y = +$.

$$\begin{array}{rcl} 2d_y & = & |a - p_2| + |b - p_2| + |(1 - a) - (1 - p_2)| \\ & & + |(1 - b) - (1 - p_2)| \\ & = & |a - b| + |a - b| \\ & = & 2r \end{array}$$

Hence, in particular, we may write $d_y = a - b$.

Observe further that $p_2$, the minority probability of $G$ when $Y = +$, is a weighted average of $a$ and $b$, namely, $p_2 = p_1 a + q_1 b$. Combining this with the equation $d_y = a - b$ allows us to express $a$ and $b$ in terms of the remaining variables, to wit: $a = p_2 + q_1 d_y$ and $b = p_2 - p_1 d_y$.

In order to prove theorem 3, we need to show that the area of disagreement $(B \cup C)$ upper bounds the area of the minority value of $F$ $(A \cup B)$. This is true just in case $C$ is larger than $A$, which is to say, if $bq_1 \geq ap_1$. Substituting our expressions for $a$ and $b$ into this inequality and solving for $d_y$ yields:

$$d_y \leq p_2 \frac{q_1 - p_1}{2p_1 q_1}$$

In short, disagreement upper bounds the minority probability just in case weak rule dependence is satisfied, proving the theorem.

## 8 The Greedy Agreement Algorithm

Dasgupta, Littman, and McAllester suggest a possible algorithm at the end of their paper, but they give only the briefest suggestion, and do not implement or evaluate it. I give here an algorithm, the Greedy Agreement Algorithm,

```
Input: seed rules F, G
loop
   for each atomic rule H
      G' = G + H
      evaluate cost of (F,G')
      keep lowest-cost G'
   if G' is worse than G, quit
   swap F, G'
```

Figure 4: The Greedy Agreement Algorithm

that constructs paired rules that agree on unlabeled data, and I examine its performance.

The algorithm is given in figure 4. It begins with two seed rules, one for each view. At each iteration, each possible extension to one of the rules is considered and scored. The best one is kept, and attention shifts to the other rule.

A complex rule (or classifier) is a list of atomic rules $H$, each associating a single feature $h$ with a label $\ell$. $H(x) = \ell$ if $x$ has feature $h$, and $H(x) = \perp$ otherwise. A given atomic rule is permitted to appear multiple times in the list. Each atomic rule occurrence gets one vote, and the classifier's prediction is the label that receives the most votes. In case of a tie, there is no prediction.

The cost of a classifier pair $(F, G)$ is based on a more general version of theorem 2, that admits abstaining rules. The following theorem is based on (Dasgupta et al., 2001).

**Theorem 4** *If view independence is satisfied, and if $F$ and $G$ are rules based on different views, then one of the following holds:*

$$\begin{array}{rcl} \Pr[F \neq Y | F \neq \perp] & \leq & \frac{\delta}{\mu - \delta} \\ \Pr[\bar{F} \neq Y | \bar{F} \neq \perp] & \leq & \frac{\delta}{\mu - \delta} \end{array}$$

*where $\delta = \Pr[F \neq G | F, G \neq \perp]$, and $\mu = \min_u \Pr[F = u | F \neq \perp]$.*

In other words, for a given binary rule $F$, a pessimistic estimate of the number of errors made by $F$ is $\delta/(\mu - \delta)$ times the number of instances labeled by $F$, plus the number of instances left unlabeled by $F$. Finally, we note that the cost of $F$ is sensitive to the choice of $G$, but the cost of $F$ with respect to $G$ is not necessarily the same as the cost of $G$ with respect to $F$. To get an overall cost, we average the cost of $F$ with respect to $G$ and $G$ with respect to $F$.

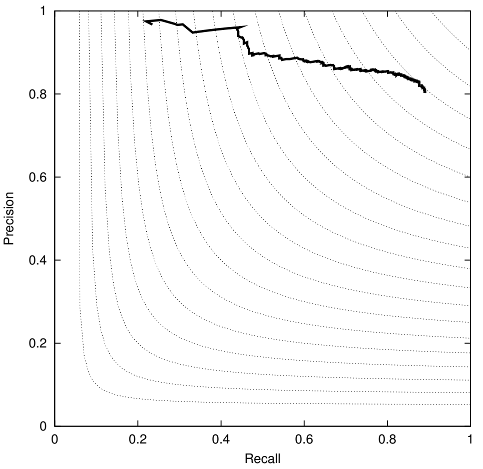


Figure 5: Performance of the greedy agreement algorithm

Figure 5 shows the performance of the greedy agreement algorithm after each iteration. Because not all test instances are labeled (some are neither persons nor locations nor organizations), and because classifiers do not label all instances, we show precision and recall rather than a single error rate. The contour lines show levels of the F-measure (the harmonic mean of precision and recall). The algorithm is run to convergence, that is, until no atomic rule can be found that decreases cost. It is interesting to note that there is no significant overtraining with respect to F-measure. The final values are 89.2/80.4/84.5 recall/precision/F-measure, which compare favorably with the performance of the Yarowsky algorithm (83.3/84.6/84.0). (Collins and Singer, 1999) add a special final round to boost recall, yielding 91.2/80.0/85.2 for the Yarowsky algorithm and 91.3/80.1/85.3 for their version of the original co-training algorithm. All four algorithms essentially perform equally well; the advantage of the greedy agreement algorithm is that we have an explanation for *why* it performs well.

## 9 The Yarowsky Algorithm

For Yarowsky's algorithm, a classifier again consists of a list of atomic rules. The prediction of the classifier is the prediction of the first rule in the list that applies. The algorithm constructs a classifier iteratively, beginning with a seed rule. In the variant we consider here, one atomic rule is added at each iteration. An atomic rule $F_\ell$ is chosen only if its precision, $\Pr[G_\ell = + | F_\ell = +]$ (as measured using the labels assigned by the current classifier $G$), exceeds a fixed threshold $\theta$.[1]

Yarowsky does not give an explicit justification for the algorithm. I show here that the algorithm can be justified on the basis of two independence assumptions. In what follows, $F$ represents an atomic rule under consideration, and $G$ represents the current classifier. Recall that $Y_\ell$ is the set of instances whose true label is $\ell$, and $G_\ell$ is the set of instances $\{x : G(x) = \ell\}$. We write $G_*$ for the set of instances labeled by the current classifier, that is, $\{x : G(x) \neq \perp\}$.

The first assumption is *precision independence.*

**Definition 5** *Candidate rule $F_\ell$ and classifier $G$ satisfy* precision independence *just in case*

$$P(Y_\ell|F_\ell, G_*) = P(Y_\ell|F_\ell)$$

*A bootstrapping problem instance satisfies precision independence just in case all rules $G$ and all atomic rules $F_\ell$ that nontrivially overlap with $G$ (both $F_\ell \cap G_*$ and $F_\ell - G_*$ are nonempty) satisfy precision independence.*

Precision independence is stated here so that it looks like a conditional independence assumption, to emphasize the similarity to the analysis of co-training. In fact, it is only "half" an independence assumption—for precision independence, it is *not* necessary that $P(Y_\ell|\bar{F}_\ell, G_*) = P(Y_\ell|\bar{F}_\ell)$.

The second assumption is that classifiers make balanced errors. That is:

$$P(Y_\ell, G_{\bar{\ell}}|F_\ell) = P(Y_{\bar{\ell}}, G_\ell|F_\ell)$$

Let us first consider a concrete (but hypothetical) example. Suppose the initial classifier correctly labels 100 out of 1000 instances, and makes no mistakes. Then the initial precision is 1 and recall is 0.1. Suppose further that we add an atomic rule that correctly labels 19 new instances, and incorrectly labels one new instance. The rule's precision is 0.95. The precision of the new classifier (the old classifier plus the new atomic rule) is 119/120 = 0.99. Note that the new precision lies between the old precision and the precision of the rule. We will show that this is always the case, given precision independence and balanced errors.

We need to consider several quantities: the precision of the current classifier, $P(Y_\ell|G_\ell)$; the precision of the rule under consideration, $P(Y_\ell|F_\ell)$; the precision of the rule on the current labeled set, $P(Y_\ell|F_\ell G_*)$; and the precision of the rule as measured using estimated labels, $P(G_\ell|F_\ell G_*)$.

The assumption of balanced errors implies that measured precision equals true precision on labeled instances, as follows. (We assume here that all instances have true labels, hence that $\bar{Y}_\ell = Y_{\bar{\ell}}$.)

$$\begin{array}{rcl} P(F_\ell G_\ell \bar{Y}_\ell) & = & P(F_\ell G_{\bar{\ell}} Y_\ell) \\ P(F_\ell G_\ell Y_\ell) + P(F_\ell G_\ell \bar{Y}_\ell) & = & P(Y_\ell F_\ell G_\ell) + P(F_\ell G_{\bar{\ell}} Y_\ell) \\ P(F_\ell G_\ell) & = & P(Y_\ell F_\ell G_*) \\ P(G_\ell|F_\ell G_*) & = & P(Y_\ell|F_\ell G_*) \end{array}$$

This, combined with precision independence, implies that the precision of $F_\ell$ as measured on the labeled set is equal to its true precision $P(Y_\ell|F_\ell)$.

Now consider the precision of the old and new classifiers at predicting $\ell$. Of the instances that the old classifier labels $\ell$, let $A$ be the number that are correctly labeled and $B$ be the number that are incorrectly labeled. Defining $N_t = A + B$, the precision of the old classifier is $Q_t = A/N_t$. Let $\Delta A$ be the number of new instances that the rule under consideration correctly labels, and let $\Delta B$ be the number that it incorrectly labels. Defining $n = \Delta A + \Delta B$, the precision of the rule is $q = \Delta A/n$. The precision of the new classifier is $Q_{t+1} = (A + \Delta A)/N_{t+1}$, which can be written as:

$$Q_{t+1} = \frac{N_t}{N_{t+1}} Q_t + \frac{n}{N_{t+1}} q$$

That is, the precision of the new classifier is a weighted average of the precision of the old classifier and the precision of the new rule.

[1] (Yarowsky, 1995), citing (Yarowsky, 1994), actually uses a superficially different score that is, however, a monotone transform of precision, hence equivalent to precision, since it is used only for sorting.

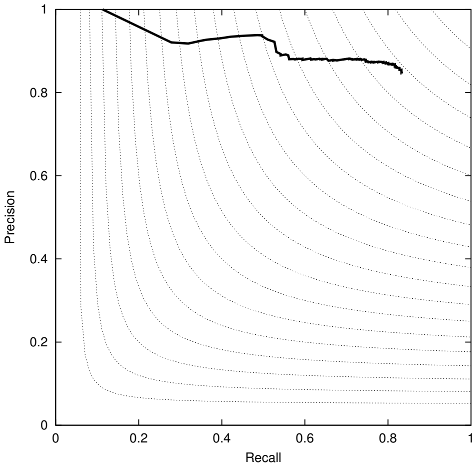


Figure 6: Performance of the Yarowsky algorithm

An immediate consequence is that, if we only accept rules whose precision exceeds a given threshold $\theta$, then the precision of the new classifier exceeds $\theta$. Since measured precision equals true precision under our previous assumptions, it follows that the true precision of the final classifier exceeds $\theta$ if the measured precision of every accepted rule exceeds $\theta$.

Moreover, observe that recall can be written as:

$$\frac{A}{N_\ell} = \frac{N_t}{N_\ell} Q_t$$

where $N_\ell$ is the number of instances whose true label is $\ell$. If $Q_t > \theta$, then recall is bounded below by $N_t\theta/N_\ell$, which grows as $N_t$ grows.

Hence we have proven the following theorem.

**Theorem 5** *If the assumptions of precision independence and balanced errors are satisfied, then the Yarowsky algorithm with threshold $\theta$ obtains a final classifier whose precision is at least $\theta$. Moreover, recall is bounded below by $N_t\theta/N_\ell$, a quantity which increases at each round.*

Intuitively, the Yarowsky algorithm increases recall while holding precision above a threshold that represents the desired precision of the final classifier. The empirical behavior of the algorithm, as shown in figure 6, is in accordance with this analysis.

We have seen, then, that the Yarowsky algorithm, like the co-training algorithm, can be justified on the basis of an independence assumption, precision independence. It is important to note, however, that the Yarowsky algorithm is not a special case of co-training. Precision independence and view independence are distinct assumptions; neither implies the other.[2]

## 10 Conclusion

To sum up, we have refined previous work on the analysis of co-training, and given a new co-training algorithm that is theoretically justified and has good empirical performance.

We have also given a theoretical analysis of the Yarowsky algorithm for the first time, and shown that it can be justified by an independence assumption that is quite distinct from the independence assumption that co-training is based on.

## References

A. Blum and T. Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann Publishers.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *EMNLP*.

Sanjoy Dasgupta, Michael Littman, and David McAllester. 2001. PAC generalization bounds for co-training. In *Proceedings of NIPS*.

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution. In *Proceedings ACL 32*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.

[2] To see that view independence does not imply precision indepence, consider an example in which $G = Y$ always. This is compatible with rule independence, but it implies that $P(Y|FG) = 1$ and $P(Y|F\bar{G}) = 0$, violating precision independence.